ISSN 2228-9860 eISSN 1906-9642 CODEN: ITJEA8

International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies

http://TuEngr.com

AI-Feature: A Software for Automatic Extraction of Features Attributes from Social Medial Texts Using AI Services

Fahim K Sufi[1*], Musleh AlSulami[2*]

[1] *Federal Government, Melbourne, AUSTRALIA.*
[2] *Umm Al-Qura University, Makkah, Makkah, SAUDI ARABIA.*
**Corresponding author (Email: research@fahimsufi.com, mhsulami@uqu.edu.sa).*

Paper ID: 13A4F

Volume 13 Issue 4

Received 23 November 2021
Received in revised form 18 March 2022
Accepted 25 March 2022
Available online 01 April 2022

Keywords:
Automated Feature Extraction; Sentiment Analysis; Named Entity Recognition; Category Classification; Feature Attribute Generation; Preprocessing for AI Algorithm.

Abstract

AI-Feature uses artificial intelligence (AI) based services like entity detection, sentiment analysis and category classification to generate an elaborate set of feature attributes that can be further analyzed by Artificial Intelligence (AI) or Machine Learning (ML) based algorithms for producing deep intelligence and insights. A data scientist, data engineer or data analyst can use AI-Feature for pre-processing text input from social media, websites, blog posts or any number of media messages, on which AI-based algorithms can operate seamlessly. AI-Feature makes the task of feature attribute extraction fully automated with Microsoft Power Automate.

Disciplinary: Artificial Intelligence

Cite This Article:

Sufi, F.K. AlSulami, M. (2022). AI-Feature: A Software for Automatic Extraction of Features Attributes from Social Medial Texts Using AI Services. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies, 13*(4), 13A4F, 1-8. http://TUENGR.COM/V13/13A4F.pdf DOI: 10.14456/ITJEMAST.2022.69

1 Background & Motivation

AI and ML algorithms have traditionally been used for identifying the correlations between a range of feature attributes with respect to an outcome variable [1] [2] [3] [4] [5] [6]. In [1], AI-based Algorithms like automated linear regression, logistic regression, anomaly detection and decomposition analysis identified and explained the correlations between various landslide feature attributes with landslide casualty. Feature attributes from biological signals were used with ML-based clustering algorithms like expectation-maximization in [2] [3], to identify cardiac abnormalities. In [4], feature attributes present in Electrocardiogram (ECG) signal was used to

identify a person with ML-based algorithms. Hence, in all our previous works [1] [2] [3] [4], feature attributes were mandated to be present before applying any ML or AI-based methods. This introduces a new problem of what happens when feature attributes are not readily available within the input data and the researcher anticipates harnessing the power of AI-based methods.

This paper proposes a method called AI-Feature that can automatically extract a range of feature attributes from the textual information, which can then be used for ML bases methods to train or predict an outcome. This method can readily be used as a preprocessor before running AI-based deep intelligence routines on social media monitoring sites (like Facebook, Instagram, Twitter), online news agencies (like BBC, CNN, NY Times, Reuter), and Government or Military websites.

AI-Feature uses a range of AI-based services and algorithms like entity detection, sentiment analysis, and category classification for automatic extraction of the feature attributes. As seen in Figure 1, a text message from a popular social media platform like Twitter mentions "Police and Tradies clashes in Melbourne amid COVID-19 crisis". This Tweet text comes with a very limited set of feature attributes like Tweet ID, Date/Time of Tweet, User, etc. Therefore, if we wanted to exploit AI algorithms like clustering or regression on this message, we won't be able to extract any meaningful insights. AI-Feature can work on this tweet text and extract a range of feature attributes automatically like {{Active Group=Police}, {Passive Group=Tradies}, {Location=Melbourne}, {Situation=COVID-19}, {Sentiment=Negative}, {Positive sentiment confidence=0.05}, {Negative sentiment confidence=0.75}, {Neutral Sentiment confidence=0.20}, {Category Classification=Issues}}. Therefore, with these wide ranges of feature attributes that are extracted by AI-Feature, clustering or prediction algorithms can generate significant insights [1].

Figure 1: Schematic diagram of AI-Feature, which uses entity detection, sentiment analysis and category classification to extract a range of feature attributes for further processing by ML models

2 Software Algorithm Overview

AI-Feature essentially combines three different techniques like entity detection, sentiment analysis and category classification to generate a wide range of feature attributes. Since the extracted feature attributes are meant to be used by ML-based classification or prediction algorithms, the AI-Feature method is effectively a pre-processor to further ML-based analysis. The individual components of AI-Feature, namely Entity detection, sentiment analysis and category classification, already have their established use cases. However, AI-Feature combines entity detection, sentiment analysis and category classification to perform a completely different task,

which is preprocessing the inputs for further analysis by clustering and prediction algorithms. Figure 2 shows the step-by-step process for AI-Feature to pre-process a series of text messages. For all the messages, AI-Feature first applies the Detect_Entities() function that releases a series of the entity along with the corresponding entity category. Then, Detect_Sentitment() is used for obtaining sentiment (which can be negative, positive, mixed, or neutral), negative sentiment confidence, positive sentiment confidence and neutral sentiment confidence. Lastly, Detect_Classification() extracts the classification of the message (like issue, document, etc.). The entire feature attribute set thus becomes available by combining the outputs of Detect_Entities(), Detect_Sentitment() and Detect_Classification(). Details of the method will be discussed.

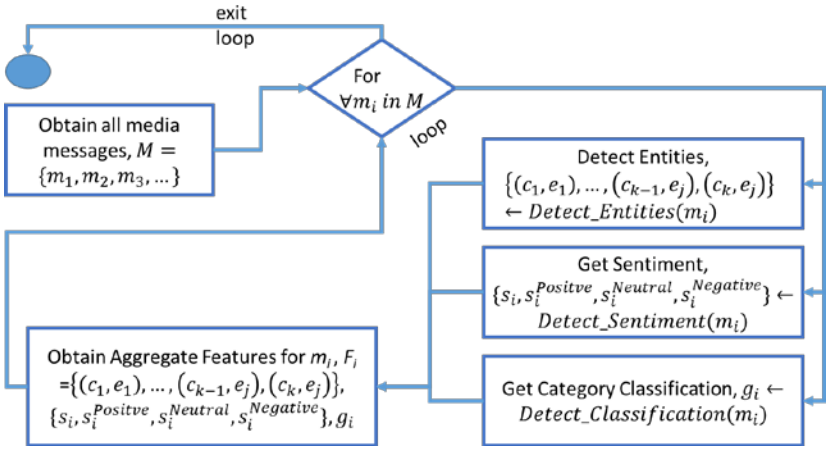

Figure 2: Step by Step working process of AI-Feature.

3 Algorithm Description

As depicted in Figure 2, the Entity Detection process returns a list of entities along with their corresponding categories as a list of lists, $\{\{c_1, e_1^{m_i}\}, \{c_2, e_2^{m_i}\}, \{c_3, e_3^{m_i}\}, \dots, \{c_k, e_i^{m_i}\}\}$. Next, the Sentiment Analysis step returns four different values for each of the message texts, as a list, $\{s_i, s_i^{Positive}, s_i^{Negative}, s_i^{Neutral}\}$, when s_i can take any sentiment category values within {positive, negative, neutral, mixed}. Moreover, $s_i^{Positive}$ or $s_i^{Negative}$ or $s_i^{Neutral}$ can take any values between 0 to 1 representing the confidence levels for each sentiment category. Then, category classification returns the scalar value of g_i. Finally, for each of the text messages, the combined feature is obtained with, $\{\{\{c_k, e_i^{m_i}\}\}, \{s_i, s_i^{Positive}, s_i^{Negative}, s_i^{Neutral}\}, g_i\}$. Algorithm 1 demonstrates the step-by-step pseudocode for extracting feature attributes with AI-Feature by using entity detection, sentiment analysis and category classification.

Algorithm 1: AI-Feature extracts feature attributes from plain text messages

Input: All the news descriptions, $M = \{m_1, m_2, m_3, \dots\}$

Output: Complete feature attribute set, $\{\{\{\{c_k, e_i^{m_i}\}\}, \{s_i, s_i^{Positive}, s_i^{Negative}, s_i^{Neutral}\}, g_i\}, \dots\}$

For $\forall$ $m_i \in M$

$\{\{c_k, e_i^{m_i}\}\} \leftarrow EntityDetection(m_i)$

$\{s_i, s_i^{Positive}, s_i^{Negative}, s_i^{Neutral}\} \leftarrow DetectSentiment(m_i)$

$g_i \leftarrow DetectClassification(m_i)$

End Loop

Entity Detection: Entity Detection is an information extraction task that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, and monetary values, percentages, etc. Entity detection has been used in almost all domains to extract key information from unstructured texts [7], [8]. Previous research has extracted 3 different types of entities (i.e., "Disease or Syndrome", "Sign or Symptom" and "Pharmacologic Substance") from health-related tweets [7] for discovering public health information and developing real-time prediction systems with respect to disease outbreak prediction and drug interactions. In [8], Basic natural language processing approaches are used to extract entities and relationships, and to identify sentiment. The keywords searched within [8] were Drug Abuse - Cannabinoids, Buprenorphine, Opioids, Sedatives and Stimulants. A study [9], qualitatively analyzed posts about methylphenidate from five French patient web forums including an analysis of information about misuse or abuse. Data were collected from French social networks that mentioned methylphenidate keywords. Text mining methods such as named entity recognition and topic modeling were used to analyze the chatter, including the identification of adverse reactions. Previous studies in [7], [8], [9] did not use entity detection as a pre-processor for AI-based algorithms. The entity Detection method could be invoked through Microsoft Power Automate [10] as seen in Figure 3.

Sentiment Analysis: Research on sentiment analysis of English text started in 2002 with the publication of two studies: [11] and [12]. A study in [11] presented a supervised learning corpus-based machine classifier and [12] presented an unsupervised classifier based on linguistic analysis. Previously, the focus of sentiment analysis was mostly on product and movie reviews. It expanded to other domains with the emergence of social media websites. Several studies followed, such as studies in [8] [9] [13]. Recent research on Sentiment Analysis has been used for assessing customer feedback towards understanding the political sentiment of people, specifically to predict election results [14]. Previous studies in [8] [9] [13] [14] did not utilize sentiment analysis as a pre-processor for AI-based algorithms to extract feature attributes from text input. The sentiment analysis method could be invoked through Microsoft Power Automate [10] as seen in Figure 3.

Category Classification*: Category classification is used to classify text inputs into categories that are useful for a specific business scenario. There are prebuild models of category classification (like Customer Feedback) available via Microsoft's AI-Builder [15] that can categorize a text input into any of the following categories:*

- Issues
- Compliment
- Customer Service
- Documentation
- Price & Billing
- Staff

Other than using assessment of customer feedback, category classification has also been used to find out the interest of an online social media user [16]. AI-Feature uses category classification along with sentiment detection and entity detection to perform feature attribute extraction for further study with the AI Algorithm. The category classification method could be invoked through Microsoft Power Automate [10], see Figure 3. AI-Feature repurposed the traditional objective of entity detection, sentiment analysis and category classification.

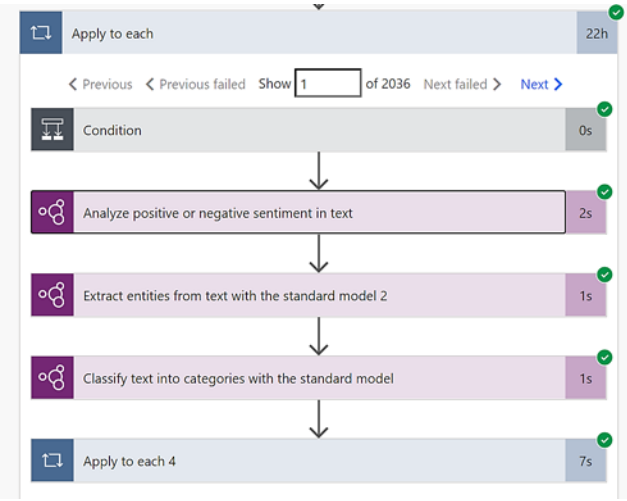

Figure 3: Implementation and evaluation of AI-Feature with Microsoft Power Automate.

4 Software Demonstration with Case Studies

We used Microsoft Power Automate to implement AI-Feature [10]. During this implementation, 2036 Tweet Messages containing the keyword "Landslide" were successfully processed with Microsoft AI builder as seen in Figure 3. During this experimentation, AI-Feature identified 2449 Organizations, 2036 category classifications, 2036 sentiments, 2036 negative sentiment confidences, 2036 neutral sentiment confidences, 2036 positive sentiment confidences, 1434 URLs, 1188 Numbers, 757 Country Regions, 682 Cities, 582 Person Names, 553 Date-Times, 308 Products, 242 Languages, 117 States, 100 Durations and a range of other attributes from tweet messages. Figure 4 shows the details of the detected attributes. As seen from Figure 4, in total there were about 18725 feature attribute values assigned to 26 different types of feature attributes during our method evaluation phase. Table I shows an example where 18 feature attributes were extracted from a tweet Message with Tweet ID '1445321702053871620'. Figure 5 shows extracted location-related entities (i.e., city, country, state, street address) being projected in Microsoft Bing Map. Hence, Figure 5 demonstrates a typical use case of AI-Feature being used to generate deep insight by automatically extracting feature attributes like locations. The entire result set acquired by running the Microsoft Power Automate package (as shown in Figure 3) along with the extracted feature attributes (i.e., in .csv format) are available at the authors' GitHub site [17].

Unlike our previous work in [1] (where landslide feature attributes were manually collated by NASA), this work features fully automated extraction of landslide feature attributes with an innovative method called AI-Feature. These feature attributes can easily be consumed in AI-based regression, anomaly detection or decomposition analysis with the techniques demonstrated in [1].

Table 1: Example Case: 17 feature attributes extracted from Twitter ID 1445321702053871620

1445321702053871620	Twitter ID
Vietnam	Country Region
21/10/2020	Date Time
13	Number
17	Number
13	Number
HoroscopeOfUSA	Organization
MercuryRetrograde	Organization
Rao Trang	Organization
RT	Organization
Thua Thien Hue	Organization
Linfa	Person Name
https://t.co/6kNuy7YB9z	URL
https://t.co/owUeLHkGxC	URL
negative	Sentiment
0.98	Negative Confidence
0	Positive Confidence
0.02	Neutral Confidence
issues	Category

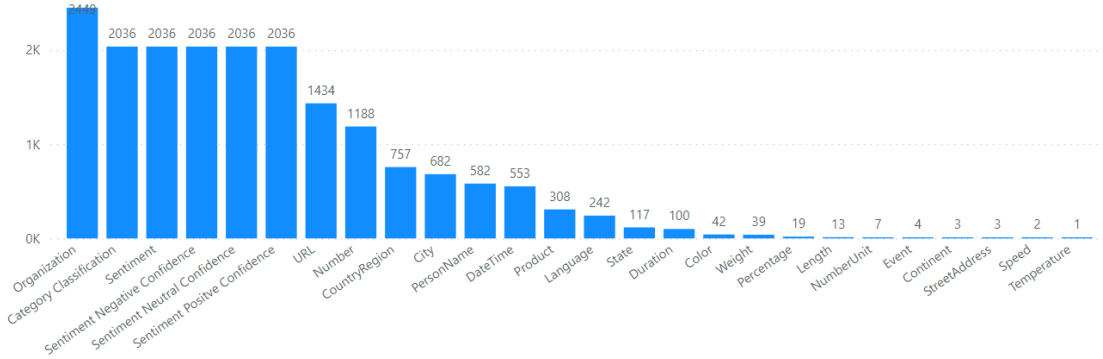

Figure 4: Number of feature attributes generated by AI-Feature against each attribute type from 2036 tweet messages.

Figure 5: Location entities extracted by AI-Feature from the landslide related tweets showing the possible geographic locations of landslides (This type of map can show the real-time occurrence of landslides by automatically extracting location entities from landslide related tweets)

5 Conclusion

As demonstrated in [7], [8], [9], entity detection was utilized as the core method for obtaining the results. Similarly, in [11] [12] [13] [14], the fundamental algorithm was sentiment analysis for acquiring the final results. In [16], category classification was executed on comprehending the interest of online users. All the existing studies demonstrated the use of sentiment analysis, entity detection and category classification as the fundamental algorithm within the respective studies to answer corresponding research questions [7] [8] [9] [11] [12] [13] [14] [16].

However, in this paper as well as in our recent studies [5] [17] [18] [19], AI-Feature uses sentiment analysis, entity detection and category classification as a predecessor to another set of AI or ML-based algorithms. The purpose of AI-Feature is to generate feature attributes for further analysis by ML-based clustering, classification, or prediction algorithms. AI-Feature works on plain texts, where seemingly no features are present for deep learning. Hence, with Algorithm 1, AI-Feature is used as a pre-processing algorithm for further deep analysis of textual information. As seen in Figure 6, without the use of AI-Feature, ML-based clustering or prediction algorithm cannot execute directly on textual information.

Following are some selected studies, where AI-Feature was successfully used:

In [5], AI-Feature automatically extracted features from 22,425 major global events from 192 countries and provided an in-depth analysis

In [18], AI-Feature automatically extracted features from Tweets containing information on 10 different types of natural disasters and generated 67528 records with 16 fields

In [19], AI-Features automatically extracted features from 1866 Tweets containing COVID-19 related information and detected 5016 entities with sentiments

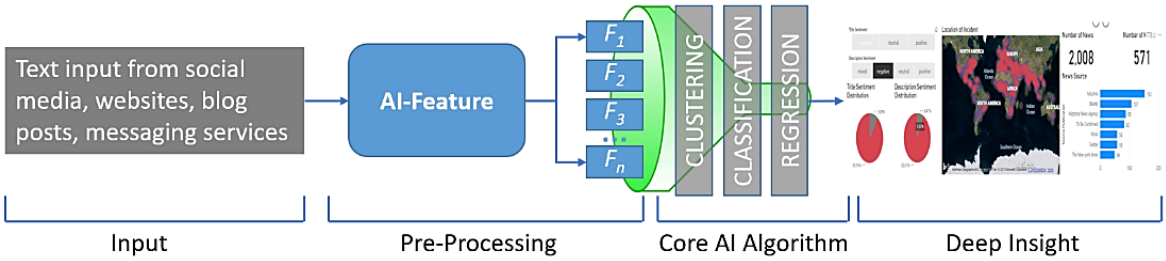

Figure 6: AI-Feature is a preprocessor to further analysis by other core AI Algorithms for producing deep insight from text inputs.

6 Availability of Data and Material

Data can be made available by contacting the corresponding author.

7 Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8 References

[1] F. K. Sufi and M. Alsulami, "Knowledge Discovery of Global Landslides Using Automated Machine Learning Algorithms," *IEEE Access*, vol. 9, pp. 131400 - 131419, 2021.

[2] F. Sufi and I. Khalil, "Diagnosis of cardiovascular abnormalities from compressed ECG: a data mining-based approach," *IEEE transactions on information technology in biomedicine*, vol. 15, no. 1, pp. 33-39, 2010.

[3] F. Sufi and I. Khalil, "A clustering based system for instant detection of cardiac abnormalities from compressed ECG," *Expert Systems with Applications*, vol. 38, no. 5, pp. 4705-4713, 2011.

[4] F. Sufi and I. Khalil, "Faster person identification using compressed ECG in time critical wireless telecardiology applications," *Journal of Network and Computer Applications*, vol. 34, no. 1, pp. 282-293, 2011.

[5] F. K. Sufi and M. Alsulami, "Automated Multidimensional Analysis of Global Events with Entity Detection, Sentiment Analysis and Anomaly Detection," *IEEE Access*, vol. 9, pp. 152449-152460, November 2021.

[6] F. K. Sufi, "AI-Landslide: Software for acquiring hidden insights from global landslide data using Artificial Intelligence," *Software Impacts*, vol. 10, no. 100177, 2001.

[7] E. Batbaatar and K. H. Ryu, "Ontology-Based Healthcare Named Entity Recognition from Twitter Messages Using a Recurrent Neural Network Approach," *International Journal of Environmental Research and Public Health*, vol. 16, no. 3628, 2019.

[8] D. Cameron, G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins and R. Falck, "PREDOSE: A Semantic Web Platform for Drug Abuse Epidemiology using Social Media," *J Biomed Inform*, vol. 46, no. 6, 2013.

[9] X. Chen, et al., A. Guenegou-Arnoux, S. Katsahian, C. Bousquet and A. Burgun, "Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate," *Frontiers in Pharmacology*, vol. 9, no. 541, 2018.

[10] Microsoft Documentation, "Microsoft Power Automate," 2021. https://docs.microsoft.com/en-us/power-automate/. (Accessed August 2021).

[11] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Conf. Empirical Methods Natural Lang. Process.*, 2002.

[12] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied," in *40th Annu. Meeting*, 2002.

[13] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter," *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS*, 2020.

[14] M. Ebrahimi, A. H. Yazdavar and A. Sheth, "Challenges of Sentiment Analysis for Dynamic Events," *IEEE Intelligent Systems*, vol. 32, no. 5, 2017.

[15] M. Documentation, "Category Classification Model," 2021. https://docs.microsoft.com/en-us/ai-builder/prebuilt-category-classification (Accessed Oct 2021).

[16] T. Hong, J.-A. Choi, K. Lim and P. Kim, "Enhancing Personalized Ads Using Interest Category Classification of SNS Users Based on Deep Neural Networks," *Sensors*, vol. 21, no. 199, 2021.

[17] F. Sufi, "AI-Feature: GitHub Source Files," 2021. https://github.com/DrSufi/DisasterAI (Accessed Oct 2021).

[18] F. K. Sufi, "Analyzing Natural Disaster related Tweets with AI and NLP based Services," *IEEE Dataport*. DOI: 10.21227/p0ed-cb23, 2021

[19] F. K. Sufi, "AI-based Automated Extraction of Entities, Entity Categories and Sentiments on COVID-19 Situation," *IEEE Dataport*, no. DOI: 10.21227/sawp-ax73, 2021

Dr.Fahim Sufi is a senior artificial intelligence solution architect with the federal government. He has held lead solution architect roles in several federal and state government agencies, including the Australian Department of Defence, the Australian Institute of Family Studies, the Victorian Department of Health, and the Victorian Department of Human Services. He obtained his PhD in computer science and information technology as well as a Master of Engineering in Computer Systems from RMIT University, Australia. His research interests include Artificial Intelligence, Machine Learning, Software Development, Big Data Analysis, Cyber, and Encryption.

Dr. Musleh Alsulami is an Assistant Professor of Information Systems at Umm Al-Qura University. He got a BSc in Software Engineering at Imam University, KSA, an MSc in Information Technology at Monash University, Australia, and a PhD in information systems at Monash University, Australia. His research interests include Enterprise Resources Planning (ERP), including ERP Life Cycles, Implementation Conflicts, Stakeholders, and Cloud-based ERP, as well as Digital Transformation in Government Organizations, Software Quality, and Human-Computer Interactions